4 - Linear Models in R

Part 3: ANOVA & ANCOVA

Eric Hare and Karsten Maurer

Iowa State University

August 22, 2013

Chick weights

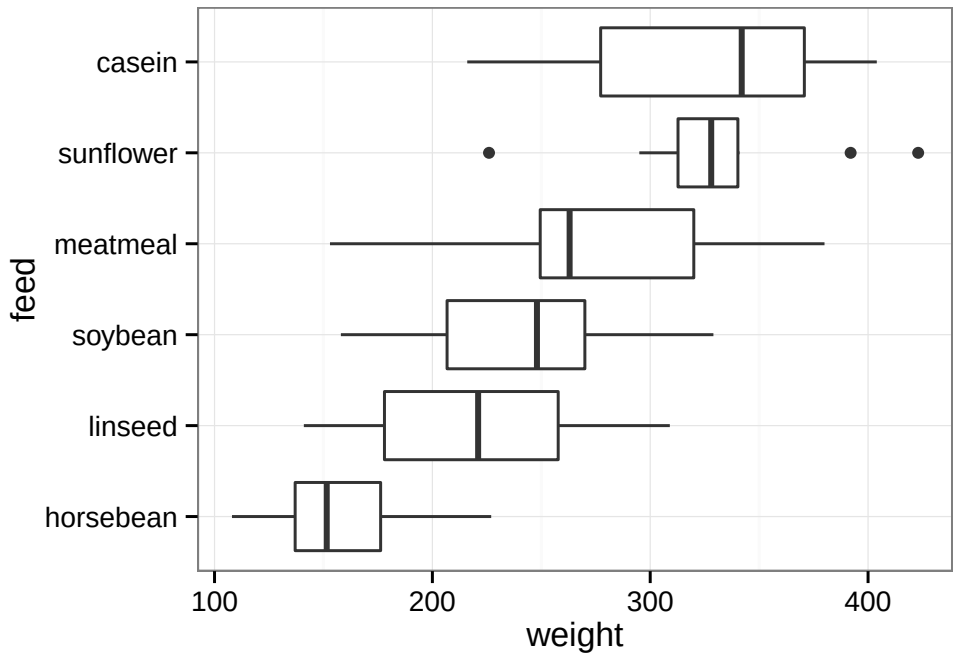
We have data on a randomized experiment examining the effect of feed supplement on the growth rate of chickens.

```
data(chickwts)
str(chickwts)
```

```
## 'data.frame': 71 obs. of  2 variables:
##  $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed  : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...
```

Chick weights

```
qplot(x = reorder(feed, weight, median), y = weight,
      data = chickwts, geom = "boxplot") +
        xlab("feed") +
        coord_flip() +
        theme_bw()
```



One-way ANOVA

ANOVA is simply a linear model with categorical predictors, so we use the same functions from regression.

```
fm1 <- lm(weight ~ feed, chickwts)
anova(fm1)

## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value  Pr(>F)
## feed       5 231129   46226    15.4 5.9e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-way ANOVA

Notice that R fits a factor effects model by default.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}; \qquad i = 1, \ldots, a$$

```
summary(fm1)$coefficients
```

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     323.583      15.83 20.4361 5.325e-30
## feedhorsebean -163.383      23.49 -6.9568 2.068e-09
## feedlinseed   -104.833      22.39 -4.6816 1.493e-05
## feedmeatmeal   -46.674      22.90 -2.0386 4.557e-02
## feedsoybean    -77.155      21.58 -3.5756 6.654e-04
## feedsunflower    5.333      22.39  0.2382 8.125e-01
```

One-way ANOVA

- Recall that for a factor effects model, constraints must be imposed on the model matrix full rank, i.e. avoid redundancy.
  - By default, R sets α_1 as the baseline level.

```
getOption("contrasts")
```

```
##         unordered           ordered
## "contr.treatment"      "contr.poly"
```

  - We can set α_a as the baseline level.

```
options(contrasts = c("contr.SAS", "contr.poly"))
```

  - We can use sum constraints, i.e. $\sum \alpha_i = 0$.

```
options(contrasts = c("contr.sum", "contr.poly"))
```

One-way ANOVA

Alternatively, we can fit a cell means model

$$Y_{ij} = \alpha_i + \varepsilon_{ij}; \qquad i = 1, \ldots, a$$

```
fm2 <- lm(weight ~ feed - 1, data = chickwts)
summary(fm2)$coefficients
```

```
##                Estimate Std. Error t value  Pr(>|t|)
## feedcasein        323.6      15.83  20.436 5.325e-30
## feedhorsebean     160.2      17.35   9.236 1.906e-13
## feedlinseed       218.8      15.83  13.815 5.185e-21
## feedmeatmeal      276.9      16.54  16.744 2.908e-25
## feedsoybean       246.4      14.66  16.810 2.357e-25
## feedsunflower     328.9      15.83  20.773 2.115e-30
```

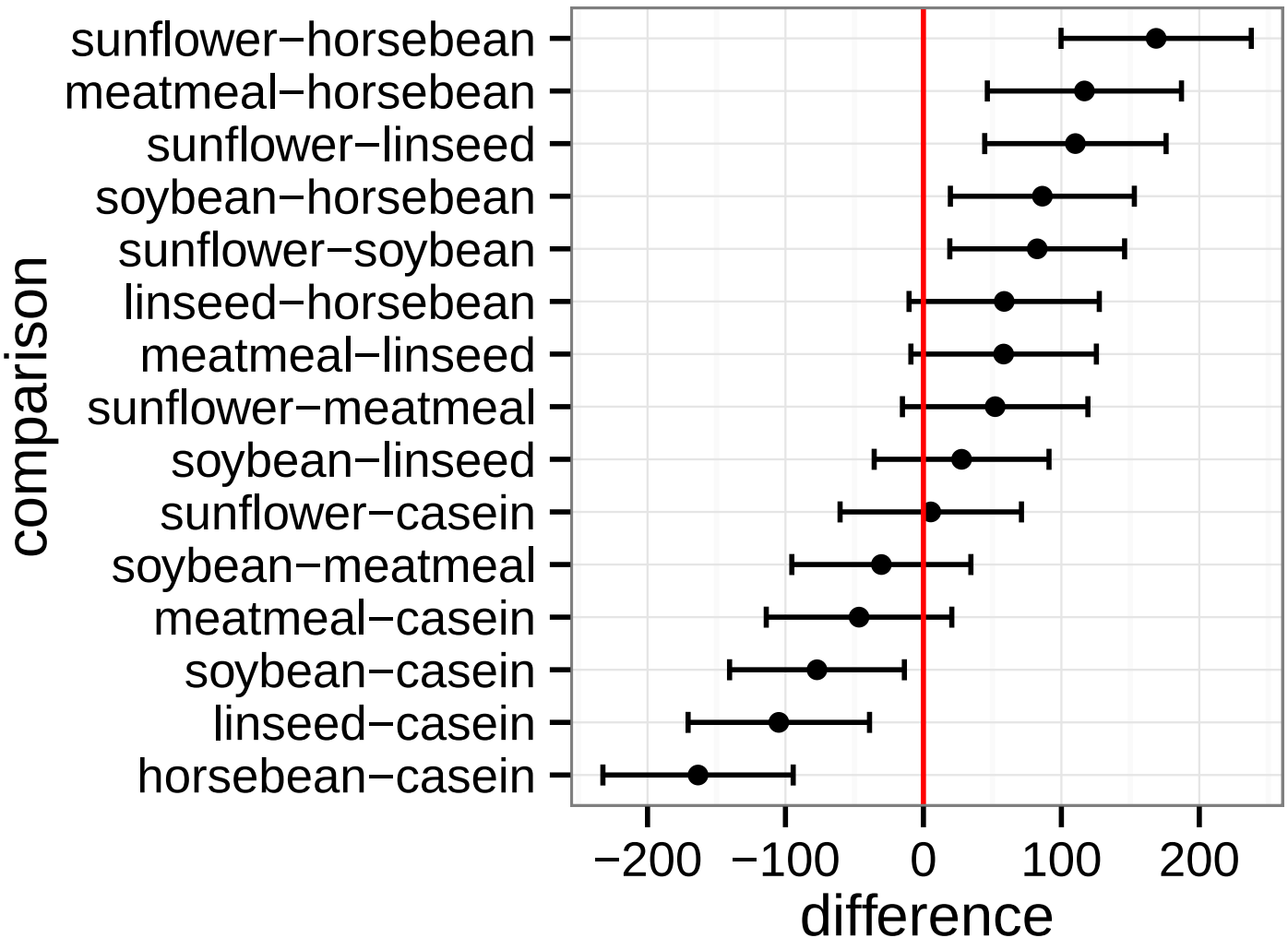
All pairwise comparisons

One easy way to adjust for all pairwise comparisons is through Tukey's honest significant difference (HSD)

```
TukeyHSD(aov(weight ~ feed, data = chickwts),
         conf.level = 0.95)$feed
```

```
##                          diff       lwr    upr     p adj
## horsebean-casein    -163.383 -232.347 -94.42 3.070e-08
## linseed-casein      -104.833 -170.587 -39.08 2.100e-04
## meatmeal-casein      -46.674 -113.906  20.56 3.325e-01
## soybean-casein       -77.155 -140.517 -13.79 8.365e-03
## sunflower-casein       5.333  -60.421  71.09 9.999e-01
## linseed-horsebean     58.550  -10.414 127.51 1.413e-01
## meatmeal-horsebean   116.709   46.335 187.08 1.062e-04
## soybean-horsebean     86.229   19.542 152.92 4.217e-03
## sunflower-horsebean  168.717   99.753 237.68 1.220e-08
## meatmeal-linseed      58.159   -9.073 125.39 1.277e-01
## soybean-linseed       27.679  -35.684  91.04 7.933e-01
## sunflower-linseed    110.167   44.413 175.92 8.843e-05
## soybean-meatmeal     -30.481  -95.375  34.41 7.391e-01
## sunflower-meatmeal    52.008  -15.224 119.24 2.207e-01
## sunflower-soybean     82.488   19.126 145.85 3.885e-03
```

All pairwise comparisons



Your turn

- Two-way ANOVA is carried out just like one-way ANOVA, but now we have two factors and a possible interaction.
- Use the `realestate` data examined in session 1, and carry out a two-way ANOVA where the response is sales price and the factors are quality and style. Before carrying out analysis, recode style so that it represents "1" or "not 1."
- Are interactions needed?
- Are both factors significant?
- Change the order of the right side of your formula. Are the results identical?

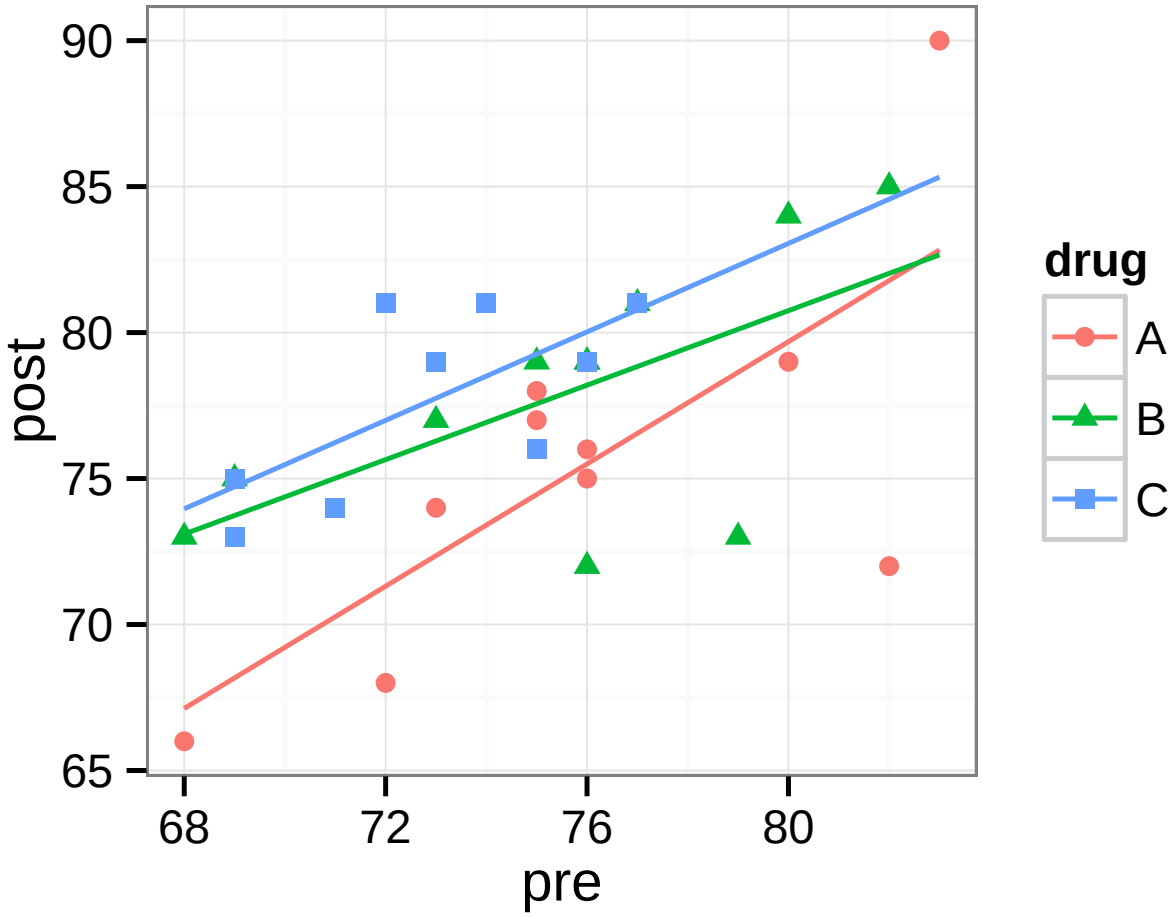
Drug side effects

- Study investigates side effects of a drug on heart function
- Treatments
  - Two forms of the drug (A, B)
  - Placebo (C)
- 30 subjects
- Heart function measured twice
  - Before administration of a treatment (PRE)
  - 2 hours after (POST)

```
heart <- read.table("data/heart.txt", header = TRUE)
str(heart)
```

```
## 'data.frame': 30 obs. of  3 variables:
##  $ drug: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ pre : int  75 68 82 76 73 83 72 75 80 76 ...
##  $ post: int  77 66 72 76 74 90 68 78 79 75 ...
```

Drug side effects



ANCOVA

- We wish use both `drug` and `pre` to describe heart function 2 hours after drug administration.

```
fm3 <- lm(post ~ pre * drug, data = heart)
anova(fm3)

## Analysis of Variance Table
##
## Response: post
##           Df Sum Sq Mean Sq F value Pr(>F)
## pre        1    242   242.1   16.88 0.0004 ***
## drug       2    100    50.2    3.50 0.0463 *
## pre:drug   2     16     8.1    0.56 0.5774
## Residuals 24    344    14.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANCOVA

```
fm4 <- update(fm3, . ~ . - pre:drug)
anova(fm4)

## Analysis of Variance Table
##
## Response: post
##           Df Sum Sq Mean Sq F value  Pr(>F)
## pre        1    242   242.1   17.47 0.00029 ***
## drug       2    100    50.2    3.62 0.04094 *
## Residuals 26    360    13.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There is evidence of a difference in heart function among the treatments (drugs).

Your turn

- `usedcars.rda` contains data on the cash offers made by 36 randomly selected used car dealers
- 12 volunteers in three age groups (young, middle, elderly) acted as owners of the same vehicle
- 6 male and 6 female volunteers were used in each age group
- offers are in hundreds of dollars
- dealer's sales volume (in hundred thousand dollars) was also recorded

Your turn

- Determine whether age of the owner impacts the dealer's offer.
- Determine whether gender of the owner impacts the dealer's offer.
- Is there an interaction between age and gender?
- If we include sales volume as a predictor, do the regression lines for each treatment have the same slope?